ways realize a given bispectrum or a bicumulant sequence, the use of a linear model as an approximation in certain applications can be justified if the computed bispectrum has an index value close to unity.

References

[1] A. M. Tekalp and A. T. Erdem, "Higher order spectrum factorization in one and two dimensions with applications in signal modeling and nonminimum phase system identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 10, pp. 1537-1549, Oct. 1989.

[2] C. L. Nikias and M. R. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. IEEE*, vol. 75, no. 7, pp. 869-891, July 1987.

[3] F. Sakaguchi and H. Sakai, "A composite linear model generating a stationary stochastic process with given bispectrum," in *Proc. Workshop Higher Order Spectral Anal.* (Vail, CO), June 28-30, 1989, pp. 24-29.

[4] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic, 1970.

The Segmental *K*-Means Algorithm for Estimating Parameters of Hidden Markov Models

BIING-HWANG JUANG AND L. R. RABINER

Abstract—**Statistical analysis techniques using hidden Markov models have found widespread use in many problem areas. This correspondence discusses and documents a parameter estimation algorithm for data sequence modeling involving hidden Markov models. The algorithm which we call the segmental *K*-means method uses the state-optimized joint likelihood for the observation data and the underlying Markovian state sequence as the objective function for estimation. We prove the convergence of the algorithm and compare it with the traditional Baum-Welch reestimation method. We also point out the increased flexibility this algorithm offers in the general speech modeling framework.**

I. Introduction

Consider a first-order N-state Markov chain governed by an $N \times N$ state transition probability matrix $A = [a_{ij}]$ and an initial state probability vector $\pi' = [\pi_1, \pi_2, \cdots, \pi_N]$. By definition, $\Sigma_{j=1}^{N} a_{ij} = 1$ for $i = 1, 2, \cdots, N$, and $\Sigma_{j=1}^{N} \pi_j = 1$. A state sequence $s = (s_0, s_1, \cdots, s_T)$ where $s_t \in \{1, 2, \cdots, N\} = Z_N$, the state index set, is a realization of the Markov chain with probability

$$\Pr(s|A, \pi) = \pi_{s0} \prod_{t=1}^{T} a_{s_{t-1}s_t}. \tag{1}$$

Suppose s is not observed directly. The actual observation sequence $x = (x_1, x_2, \cdots, x_T)$ where $x_t \in \mathbb{R}^K$, the usual K-dimensional Euclidean space, is a manifestation of some state sequence s through an observation probability density set $B = \{b_i\}_{i=1}^{N}$ where each b_i is defined on $\mathbb{R}^K$. For each $s_t = i$, the probability of occurrence of x_t is given by b_i. The triple $\lambda = (\pi, A, B)$ is called a hidden Markov model [4] and the density function of x is given by

$$f(x|\lambda) = \sum_{s} \pi_{s0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t). \tag{2}$$

The objective in maximum likelihood estimation is to maximize $f(x|\lambda)$ over all parameters λ for a given observation sequence x (or sequence set $X = \{x^l\}$).

The above maximum likelihood estimation problem can be effectively solved using a reestimation algorithm [5]-[7], often called the Baum-Welch algorithm [5] or forward-backward algorithm. The algorithm is an iterative procedure that guarantees a monotonic increase in the likelihood through a set of reestimation transformations.

In this correspondence, we consider a different optimization criterion for estimating the parameters of the hidden Markov model. Instead of the likelihood function (2), we use

$$\max_{s} f(x, s|\lambda) = \max_{s} \pi_{s0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t) \tag{3}$$

as the optimization objective. Note that (3) focuses on the most likely state sequence as opposed to summing over all possible state sequences as in (2). We shall call (3) the state-optimized likelihood. Cast in the formulation of a maximum *a posteriori* sequential estimate, (3) was previously addressed by Jelinek ([8, appendix II]).

The motivation for using (3) as the optimization criterion is as follows. First, the summation in (2) requires that all state transition paths be considered in the likelihood calculation, thus requiring significant computation. Second, since the b_i's in the set B often vary in value over a large dynamic range, evaluation of the likelihood along every possible path will inevitably encounter numerical difficulties. Third, in speech recognition applications, modeling and decoding must both be performed on the observation data sets and the criterion of (3) appears to be quite natural for both of these tasks. Also, the most likely state sequence s accompanying the optimization procedure carries some information, such as the state duration, which is useful in many applications [9], [10].

Although the segmental K-means algorithm has been described in a previous publication [9], we shall formally define the algorithm and discuss its convergence properties in this paper. Our proof of convergence of the algorithm essentially follows the procedure outlined in Baum's original paper [5], supplemented by [6] and [7] for different types of observation densities. It, however, requires some modifications because of the nonlinear (maximization) operator involved in (3). By making use of Zangwill's global convergence theorem [11, p. 91], we separate the issue of algorithm convergence and the issue of increasing state-optimized likelihood. A similar strategy was adopted by Sabin [12] in his proof of global convergence of the generalized Lloyd vector quantizer design algorithm. The global convergence theorem is a general result, applicable to the case where the hidden state space is a space of independent variables (as formulated in the numerous examples of the EM algorithm paper by Dempster *et al.* [13]) as well as our current consideration of the space of Markovian samples. This allows us to focus on how the algorithm guarantees an increase in the state-optimized likelihood for several kinds of observation densities.

II. Global Convergence Theorem

Let Λ be an open subset of Euclidean p space $\mathbb{R}^p$. A hidden Markov model λ is a point in Λ and to each $\lambda \in \Lambda$ we have a smooth assignment $\lambda \rightarrow (\pi(\lambda), A(\lambda), B(\lambda))$. Furthermore, we assume Λ is compact and $f(x, s|\lambda)$ is continuous in Λ and differentiable in its interior so that $f(x, s|\lambda)$ is bounded above.

An algorithm T on Λ is a mapping from points of Λ to subsets of Λ. When the mapping is point to point, T is simply a transformation. We say algorithm T on Λ is closed if $\lambda \in \Lambda$, $\zeta \in \Lambda$, $\lambda_n \rightarrow \lambda$, $\zeta_n \rightarrow \zeta$ and $\zeta_n \in T(\lambda_n)$ imply that $\zeta \in T(\lambda)$. Closure is a generalization of continuity. For point-to-point mapping, continuity implies closedness.

Manuscript received March 8, 1988; revised January 31, 1990.
The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 9036898.

0096-3518/90/0900-1639$01.00 © 1990 IEEE

Let Ω be the set of fixed points of T. A function g on Λ is called an ascent function for algorithm T if 1) g: $\Lambda \rightarrow \mathcal{R}' \subset \mathcal{R}$ is continuous (e.g., $\mathcal{R}' = (0, \infty)$); 2) $g(\zeta) > g(\lambda)$ for $\zeta \in T(\lambda)$ and $\lambda \notin \Omega$; and 3) $g(\zeta) \geq g(\lambda)$ for $\zeta \in T(\lambda)$ and $\lambda \in \Omega$. This definition of ascent function is a direct opposite of the descent function defined in [12] by Sabin.

Global Convergence Theorem: Let the sequence $\{\lambda_i\}_{i=0}^{\infty}$ *be generated by an algorithm T such that* $\lambda_{i+1} \in T(\lambda_i)$, *for some* $\lambda_0 \in \Lambda$. *Let T be closed and* $\Omega \subset \Lambda$ *be the set of fixed points of T. Then i)* Ω *is closed; ii) all the accumulation points of* $\{\lambda_i\}$ *are in* Ω *and* $g(\lambda_i)$ *converges monotonically to* $g(\lambda^*)$ *for some* $\lambda^* \in \Omega$ *if g is an ascent function.*

We say a function $h(\lambda)$ is T converging if h is an ascent function for algorithm T which satisfies all the above requirements such that $\lim_i h(\lambda_i) = h(\lambda^*)$ for some $\lambda^* \in \Omega$.

III. T-Converging Likelihood Functions

The observation densities $B = \{b_i\}_{i=1}^{N}$ we consider in the hidden Markov models are of the following types: a) strictly log-concave densities such as the normal, Poisson, binomial, and gamma, etc. [5]; b) elliptically symmetric densities [6]; c) mixtures of a) and/or b) above [7]; d) mixture densities with autoregressive constraints [10]; and e) partitioned or vector quantized mixtures of the above satisfying certain conditions [10]. These densities (as the bases of likelihood functions for parameter estimation) can be shown to be T converging for some specific T's.

For example, the strictly log-concave densities we consider are $b(x|\lambda)$ where for each i and almost all x, $\log b(x|\lambda)$ is strictly concave in λ and $\lim_{|\lambda| \rightarrow \infty} \log b(x|\lambda) = -\infty$ [5]. These densities are T converging if T is chosen as T: $\lambda \rightarrow \bar{\lambda}$ where

$$\bar{\lambda} = \arg\max_{\lambda'} Q(\lambda, \lambda') = \arg\max_{\lambda'} \int b(x|\lambda) \log b(x|\lambda')\, d\mu(x) \tag{4}$$

as show by Baum *et al.* [5]. This can be easily seen by setting Baum's N-state Markov chain to a single state case. In the above, a totally finite measurable space of x is assumed with measure $\mu(\cdot)$.

Furthermore, as will be discussed below, hidden Markov model density of (2) could be used as the observation density of a certain state in another hidden Markov model and the T-converging property can be hierarchically constructed accordingly.

IV. The Segmental K-Means Algorithms

The segmental K-means algorithm, as the name implies, is an algorithm for estimating the hidden Markov model parameters by embedding the K-means method [14] in a Markov chain modeling algorithm for time-varying data sequences. The algorithm involves iteration of two fundamental steps: 1) segmentation and 2) optimization. (A similar but more restricted algorithm bearing the name "Viterbi extraction" has also been proposed by Jelinek [8].) We start from an initial model λ. The segmentation step is equivalent to a sequential decoding (or encoding, depending on the view of source coding or channel coding) procedure and can be optimally performed via a generalized Viterbi algorithm [15] which attains $\max_s f(x, s|\lambda)$ as in (3).

Given a state sequence s and the observation x, the optimization step finds a new set of model parameters $\bar{\lambda}$ so as to maximize the above state-optimized likelihood. That is

$$\bar{\lambda} = \arg\max_{\lambda} \left\{\max_{s} f(x, s|\lambda)\right\}. \tag{5}$$

Note that maximization of the state-optimized likelihood in (5) may not be straightforward. For each state i, the generalized iteration algorithm may have to be employed, depending on the choice of the observation densities which need to be T-converging.

We then replace the original model λ by the new $\bar{\lambda}$ and iterate the above two steps until the state-optimized likelihood converges within a prescribed threshold.

V. Convergence of the Segmental K-Means Algorithm

With all the results discussed in Sections II and III, it is now straightforward to show that the segmental K-means algorithm converges in terms of the state-optimized likelihood.

Zangwill's global convergence theorem is the main theorem that the proof is based on. What needs to be shown is that the algorithm T: $\lambda \rightarrow \bar{\lambda}$ according to (3) and (5) is closed and that the state-optimized likelihood $\max_s f(x, s|\lambda)$ is an ascent function for the algorithm.

There are two difficulties encountered at this point, however. First, although we have chosen the observation densities to be T converging, it is not trivial to show that the state-optimized likelihood is T-converging because after each iteration the optimal state sequence may have been changed and, therefore, the set of data presented for the estimation of model parameters in each state may be different from the previous set. Second, as mentioned before, the T-converging property of the observation density functions only guarantees fixed point convergence and it is not readily clear how the individual T convergence of the state observation densities would affect the T convergence of the overall state-optimized likelihood with a underlying Markov chain structure.

The algorithm T is closed because we assume that the function $f(x|\lambda) = \Sigma_s f(x, s|\lambda)$ and thus $\max_s f(s|\lambda)$ is continuously differentiable in λ for almost all x in a totally finite measurable space. The inclusion of the underlying Markov chain does not induce extra complications as it only produces the product form in (2) and (3).

Let s^* and $\bar{s}$ be the two optimal state sequences:

$$s^* = \arg\max_{s} f(x, s|\lambda) \tag{6.1}$$

$$\bar{s} = \arg\max_{s} f(x, s|\bar{\lambda}). \tag{6.2}$$

Then

$$\max_{s} f(x, s|\bar{\lambda}) \geq f(x, s^*|\bar{\lambda}) \tag{7.1}$$

$$= \max_{\lambda'} f(x, s^*|\lambda') \tag{7.2}$$

$$= \max_{\lambda'} \left\{\max_{s} f(x, s|\lambda')\right\} \tag{7.3}$$

$$\geq \max_{s} f(x, s|\lambda). \tag{7.4}$$

The inequality in (7.1) is strict unless $\bar{s} = s^*$ which results in $\bar{\lambda} \in T(\bar{\lambda})$ and the inequality in (7.4) is strict unless λ achieved the maximum of (7.3) or $\lambda \in T(\lambda)$.

Note that the maximization over λ' in (7.2) and (7.3) can be replaced by any converging hill-climbing algorithm such as the generalized K-means method even though it only guarantees a fixed point solution.

This completes the proof that the segmental K-means algorithm converges in Zangwill's global convergence sense.

VI. Remarks

Although we have discussed extensively the convergence properties of the segmental K-means algorithm, some specifics of the actual transformation were not given. The algorithm of (5) can be written as

$$\begin{aligned}\bar{\lambda} &= \arg\max_{\lambda} \left\{\max_{s} f(x, s|\lambda)\right\} \\ &= \arg\max_{\lambda} \left\{\max_{s} \left[\log f(x|s, \lambda) + \log f(s|\lambda)\right]\right\}. \end{aligned} \tag{8}$$

Using the previous notation of (6.1), we see that $\max_\lambda \{\log f(x|s^*, \lambda) + \log f(s^*|\lambda)\}$ consists of two terms that can be separately optimized since $\log f(s^*|\lambda)$ is a function of $A(\lambda)$ alone and $\log f(x|s^*, \lambda)$ is a function of $B(\lambda)$ above. (We shall neglect the initial probability vector π for simplicity.)

Let $s^* = (s_0^*, s_1^*, s_2^*, \cdots, s_T^*)$, $s_t^* \in Z_N$. Then

$$\log f(x \mid s^*, \lambda) = \sum_{t=1}^{T} b_{s_t^*}(x_t) \tag{9}$$

and

$$\log f(s^* \mid \lambda) = \sum_{t=1}^{T} \log a_{s_t^*\ {}_1 s_t^*} \tag{10}$$

which can be regrouped as

$$\log f(x \mid s^*, \lambda) = \sum_{i=1}^{N} \sum_{t \in T_i} \log b_i(x_t) \tag{11}$$

and

$$\log f(s^* \mid \lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t \in T_{ij}} \log a_{s_t^*\ {}_1 s_t^*} \tag{12}$$

where $T_i = \{t: s_t^* = i\}$ and $T_{ij} = \{t: s_{t-1}^* = i, s_t^* = j\}$. Since the b_i's are independent, the optimization over $B = \{b_i\}$ can be done separately for each $i = 1, 2, \cdots, N$, according to (11). This is equivalent to N separate estimations of b_i's given N sets of data $\{x_t: t \in T_i\}$. The maximization of (12) over $A = [a_{ij}]$ subject to the constraints $\Sigma_{j=1}^{N} a_{ij} = 1$ and $a_{ij} \geq 0$ for all i is essentially a set of N maximization problems

$$\max \sum_{j=1}^{N} \sum_{t \in T_{ij}} \log a_{s_{t-1}^* s_t^*}, \qquad i = 1, 2, \cdots, N \tag{13}$$

which have the solution $\bar{a}_{ij}$

$$\bar{a}_{ij} = \frac{\| T_{ij} \|}{\sum_{j=1}^{N} \| T_{ij} \|}, \qquad i, j = 1, 2, \cdots, N \tag{14}$$

where $\| \cdot \|$ denotes the cardinality.

Equations (11) and (12) provide an intuitive confirmation about the convergence of the algorithm by way of separating the optimizations over the observation densities and the transition probabilities of the Markov chain. However, unlike the original Baum-Welch algorithm which uses (2) as the modeling criterion, the solution (14) does not require probability weighting from the (unhidden) observations x.

The segmental K-means algorithm can be straightforwardly extended to the case of multiple independent observation sequences $X = \{x^i\}$. The optimization criterion for multiple independent sequences becomes

$$\max_{S} f(X, S \mid \lambda) = \max_{S} \prod_{i} f(x^i, s^i \mid \lambda) \tag{15}$$

where $S = \{s^i\}$ is the state sequence set. The multiple sequence case only incurs an extra summation over the sequence index in the formulation results of (11) and (13) above.

It is important to note that the T-convergence property can be hierarchical as mentioned above. In the above, the proof that the segmental K-means algorithm is T converging was built upon the T-convergence properties of the observation likelihood. It should be understood that T-converging hidden Markov models are equally applicable as the (sequence) observation density. For speech modeling, this thus legitimizes the use of segmental subword hidden Markov models [16], and connected word models [9], although in the latter case the ultimate Markov chain for the digits is of little use because of the assumed random nature of the unconstrained digit sequence utterances.

References

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Analysis Mach. Intel.*, vol. PAMI-5, no. 2, pp. 179-190, Mar. 1983.

[2] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pt. 1, pp. 1075-1106, Apr. 1983.

[3] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4-16, Jan. 1986.

[4] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pt. 1, pp. 1035-1074, Apr. 1983.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164-171, 1970.

[6] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729-734, Sept. 1982.

[7] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1235-1249, July 1985.

[8] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE Proc.*, vol. 64, no. 4, Apr. 1976.

[9] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental *K*-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 64, no. 3, pp. 21-40, May 1986.

[10] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, pp. 1404-1413, Dec. 1985.

[11] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.

[12] M. J. Sabin, "Global convergence and empirical consistency of the generalized Lloyd algorithm," Ph.D. dissertation, Stanford Univ., Stanford, CA, May 1984.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, ser. 39, pp. 1-38, 1977.

[14] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat., Prob.*, vol. 1, 1967, pp. 281-296.

[15] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268-278, Mar. 1973.

[16] C. H. Lee, F. K. Soong, and B. H. Juang, "A segment model based approach to speech recognition," presented at the IEEE ICASSP-88, New York, NY, Apr. 1988.

FIR Filtering by the Modified Fermat Number Transform

WEIPING LI AND ALLEN M. PETERSON

Abstract—**Right-angle circular convolution (RCC) and the modified Fermat number transform (MFNT) are introduced. It is shown that a linear convolution of two *N* point sequences can be obtained by a corresponding *N* point RCC. It is also shown that the MFNT supports RCC so that a linear convolution can be computed by an *N* point MFNT and its inverse plus *N* multiplies.**

Manuscript received April 10, 1987; revised October 23, 1989.

W. Li is with the Computer Science and Electrical Engineering Department, Lehigh University, Bethlehem, PA 18015.

A. M. Peterson is with the Electrical Engineering Department, Stanford University, Stanford, CA 94305.

IEEE Log Number 9036894.

0096-3518/90/0900-1641$01.00 © 1990 IEEE